

The Novel Object and Unusual Name (NOUN) database: a collection of novel images for use in experimental research

Article (Accepted Version)

Horst, Jessica S and Hout, Michael C (2016) The Novel Object and Unusual Name (NOUN) database: a collection of novel images for use in experimental research. Behavior Research Methods, 48 (4). pp. 1393-1409. ISSN 1554-351X

This version is available from Sussex Research Online: <http://sro.sussex.ac.uk/id/eprint/55936/>

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:

Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**The Novel Object and Unusual Name (NOUN) Database: a collection of novel images for
use in experimental research**

Jessica S. Horst^{1*} & Michael C. Hout²

¹University of Sussex, ²New Mexico State University

Address correspondence to: Jessica S. Horst, School of Psychology, Pevensey 1 Building,
Falmer, Brighton, East Sussex, BN1 9QH, United Kingdom, tel: +44 (0) 1273 873084, email:
jessica@sussex.ac.uk.

Acknowledgements

The creation of the original NOUN Database was funded by British Academy Small Grant SG-50687 to J.S.H. The expansion of the database and the preparation of this manuscript was made possible by British Academy/Leverhulme Trust Senior Research Fellowship SF120032 to J.S.H. We thank Ryan Kavlie for photographing the objects and editing the photographs in the NOUN Database. We would also like to thank Emily Scott for her assistance with some of the materials and Zoe Flack for help with data collection. We are indebted to Robert Goldstone for introducing the authors to each other.

Abstract

Many experimental research designs require images of novel objects. Here we introduce the Novel Object and Unusual Name (NOUN) Database. This database contains 64 primary novel object images and additional novel exemplars for ten basic- and nine global-level object categories. The objects' novelty was confirmed by both self-report and a lack of consensus on questions that required participants to name and identify the objects. We also found that object novelty correlated with qualifying naming responses pertaining to the objects' colors. Results from a similarity sorting task (and subsequent multidimensional scaling analysis on the similarity ratings) demonstrated that the objects are complex and distinct entities that vary along several featural dimensions beyond simply shape and color. A final experiment confirmed that additional item exemplars comprise both sub- and superordinate categories. These images may be useful in a variety of settings, particularly for developmental psychology and other research in language, categorization, perception, visual memory and related domains.

The Novel Object and Unusual Name (NOUN) Database: a collection of novel images for use in experimental research

Many psychological experiments involve word learning tasks, in which participants—both adults and children—are taught names for objects either explicitly, for instance, through the use of social cues and ostensive feedback (Gauthier & Tarr, 1997; Horst & Samuelson, 2008) or implicitly across several encounters (e.g., Axelsson & Horst, 2014; Yu & Smith, 2007). In studies such as these, novelty is critical for ensuring that researchers are testing learning that occurred as a function of the experimental manipulation and not merely tapping into knowledge acquired prior to the experiment (Ard & Beverly, 2004; Bornstein & Mash, 2010). In addition, novelty is often critical for categorization studies in which participants must learn to extrapolate information from one category exemplar and generalize or apply that information to new exemplars (e.g., Homa, Hout, Milliken, & Milliken, 2011; J. D. Smith & Minda, 2002).

Previous research demonstrates that novelty (or lack thereof) can have a profound effect on subsequent learning. For example, after toddlers explore novel objects for 1-2 minutes they are significantly less likely to associate novel names with those objects than with still-novel objects (Horst, Samuelson, Kucker, & McMurray, 2011; see also Kucker & Samuelson, 2012). Thus, even brief prior experience with stimuli can change subsequent behavior on critical test trials. Likewise, gradual, prior experience with stimuli can also influence subsequent behavior, such as looking times during the learning phase of an object-examination categorization task (Bornstein & Mash, 2010). It is therefore ideal that visual stimuli have not been seen before, in order to ensure that any inferences made regarding learning were not actually due to participants' exposure to the items prior to the experiment. For many experimental designs it is important that objects are also easy to distinguish from each other (e.g., Twomey, Ranson, & Horst, 2014; Yu

& Smith, 2007), however, for other designs it can be useful to have objects that are somewhat similar (e.g., Homa et al., 2011; Hout & Goldinger, 2015).

There are existing databases of known, familiar, real-world objects (e.g., Brady, Konkle, Alvarez, & Oliva, 2008; Dan-Glauser & Scherer, 2011; Hout, Goldinger, & Brady, 2014; Konkle, Brady, Alvarez, & Oliva, 2010; Migo, Montaldi, & Mayes, 2013) and human faces (e.g., Ebner, Riediger, & Lindenberger, 2010; Matheson & McMullen, 2011) as well as databases of related items uniquely optimal for use in categorization studies (e.g., Gauthier & Tarr, 1997; Marchewka, Żurawski, Jednoróg, & Grabowska, 2014). However, for researchers investigating memory for objects and object names, there is a critical need for a database of novel objects for use in such experiments where it is critical that participants have no *a priori* knowledge of the stimuli and that the objects are not already associated with specific names. The NOUN database is such a collection of novel object images.

Why Use the NOUN Database?

The NOUN Database offers several advantages for researchers requiring images of unusual objects. First, the images in the NOUN Database depict multi-part, multi-colored, real three-dimensional (3D) objects as opposed to simple geometric shape configurations (e.g., L. B. Smith & Yu, 2008; Wu, Gopnik, Richardson, & Kirkham, 2011) or seemingly animate objects (e.g., Gauthier & Tarr, 1997; Mather, Schafer, & Houston-Price, 2011; Rakison & Poulin-Dubois, 2002). As such, these stimuli are ideal for researchers who need images of naturalistic, complex novel objects to present against images of real 3D objects that are already familiar to participants (e.g., familiar distractors or known competitors). Indeed, complex novel objects are often presented against known objects, for example in language research (e.g., Axelsson & Horst, 2014; Giezen, Escudero, & Baker, in press; Mather & Plunkett, 2009; Warren & Duff, 2014;

Zosh, Brinster, & Halberda, 2013). In such cases it is vital that the novel stimuli are just as “credible” as the familiar, known objects, which requires similar shading, colors, textures and complexity. The stimuli in the NOUN Database have such properties because they are images of real objects (e.g., they are not “impossible” objects that might be created from a software package).

Second, researchers frequently choose their stimuli based on their own intuitive judgments of novelty and similarity (Migo et al., 2013). This practice is especially prevalent in developmental psychology, where researchers make assumptions about objects that are unlikely to be familiar to young children without prior confirmation (but see Horst & Samuelson, 2008 for a quick confirmation method). This can be problematic because children may be implicitly learning about the object categories although they have not yet heard the category names. In experiments requiring images of novel objects, this problem can be avoided by using the pre-existing NOUN Database; the novelty and similarity ratings we present can inform researchers’ decisions on which stimuli to use depending on their research questions. Specifically, these ratings can be used to ensure a subset of stimuli are equally novel and not already associated with a specific name as well as to vary the novelty or similarity across items.

Relatedly, using an existing database facilitates comparison across experiments, which can be especially helpful when different experiments address unique, but related research questions or when one wants to compare related effects. For example, young children generalize names for nonsolid substances to other substances made of the same material (Imai & Gentner, 1997; Soja, Carey, & Spelke, 1992), but only when the substances share *both* the same material and color—if the stimuli only share the same material this effect disappears (Samuelson & Horst, 2007). Similarly, adults are faster to repeat non-words from high-density lexical neighborhoods than

low-density neighborhoods (Vitevitch & Luce, 1998), but this effect disappears with different stimuli, e.g., when the leading and trailing silences in the audio files are removed to equate stimulus duration (Lipinski & Gupta, 2005). The use of existing stimuli is also consistent with the recent push in the psychology research community to share resources and to facilitate replicability (for a lengthier discussion, see Asendorpf et al., 2013).

Finally, using an existing set of stimuli saves time and reduces research expenses because data collection on the substantive experiment can begin quickly without the need for additional preliminary experiments that ensure that the stimuli are in fact novel and unlikely to be already be associated with a particular name. The current Experiment 1 is effectively a preliminary experiment conducted on behalf of researchers who wish to use the NOUN Database. This is valuable because obtaining and selecting experimental stimuli is often a highly time-consuming phase of the research process (Dan-Glauser & Scherer, 2011; see also Umla-Runge, Zimmer, Fu, & Wang, 2012). Even using 3D rendering software can take time to learn and be expensive. Moreover, using an existing database utilized by multiple researchers may expedite the ethical approval process for new studies. Taken together, these time-saving aspects make the NOUN Database particularly useful to students who must conduct research projects quickly with a strict deadline as well as early career researchers who especially benefit from reduced time-to-publication (see A. K. Smith et al., 2011, for a related argument).

Although one advantage of the NOUN Database is its ability to save valuable time and money, researchers using the database may choose to conduct their own preliminary experiments to ensure the stimuli that will be used in their main experiment are in fact novel to their participant pool. Researchers can also use the images in the NOUN Database as a supplement to their own stimuli, which offers even greater experimental flexibility. Note, a second advantage of

the NOUN Database is its size: it includes 64 items, which is many times more than is often required for most studies using images of novel objects (e.g., 2 novel objects, Bion, Borovsky, & Fernald, 2013; 3 novel objects, Rost & McMurray, 2009; 1 novel object, Werker, Cohen, Lloyd, Casasola, & Stager, 1998).

The Current Experiments

The original NOUN Database included 45 images of novel objects (Horst, 2009). Each object is distinct, and together the objects include a variety of shapes, colors and materials. We have expanded the NOUN Database to include a total of 64 objects. In Experiment 1, adult participants judged the novelty of each object. Participants were asked whether they were familiar with each object, then what they would call each object, and finally what they thought each object really was. In Experiment 2, we used a multi-dimensional scaling (MDS) task to examine the extent to which the objects were complex and distinct. Specifically, we wanted to ensure that the objects were complex enough such that participants were not only appreciating one or two featural dimensions (e.g., color and shape) when considering the objects. Finally, in Experiment 3, we repeated the MDS task with multiple exemplars from ten of the object categories to determine the relationships between the various novel object categories, enabling us to make recommendations as to which subcategories belong to the same global-level categories.

This database was originally created for use in word learning experiments, primarily with children. However, researchers may also require novel objects when investigating categorization (e.g., Twomey et al., 2014), (visual) short-term memory (e.g., Kwon, Luck, & Oakes, 2014) and long-term visual memory (e.g., Hout & Goldinger, 2010, 2012). This newly developed set of photographs is freely available to the scientific community from the authors for noncommercial use.

Experiment 1

Our goal with this first experiment was to test whether the novel objects in the NOUN Database are, in fact, generally novel across participants. Novelty is on a continuum (Horst, 2013) and could imply a stimulus has never been encountered before or has been previously encountered but never associated with a particular name. Due to this plurality of possible definitions, we examined the novelty of the NOUN stimuli using multiple tasks that each tap into a different (but related) aspect of what it means for something to be novel. First, we simply asked participants if they had seen each object before. One might argue that this is the ultimate test of novelty. Next, we asked participants what they would call each object. It may be that an object has never been seen before but is highly reminiscent of a known object. The name question allowed us to examine this conceptualization of novelty. Finally, we asked participants what they thought each object really was. This question was included (in addition to the name question, with which it may seem partially redundant) because it is possible to name something based on its appearance but to know that it is not really from a particular category. For example, someone might know that something looks like a clothespin but is really art (Landau, Smith, & Jones, 1998), looks like a bow tie but is really pasta or looks like a jiggy but is really a zimbo (Gentner, 1978).

Method

Participants. Undergraduate students participated for course credit. Each participant gave written informed consent and then completed all three experiments. Before data collection began, the authors agreed on a target sample size of $n = 30$ based on previous studies of adult ratings of stimuli for use in experiments with children (e.g., Horst & Twomey, 2013; Samuelson & Smith, 1999, 2005). Participants signed up to participate using an online sign-up system. Initially, 47

students signed up to participate before sign-ups were closed. Six students canceled and another six students failed to attend their scheduled sessions. The remaining 35 students showed up to participate. Data from three participants were excluded from all analyses because of equipment failure ($n = 1$) and failure to follow the instructions ($n = 2$). This resulted in a final sample of 32 participants (20 women). Neither author analyzed any data until after all 32 participants had completed the study.

Materials. Photographs were taken of real, 3D objects against a white background. Raw images were then imported into Adobe Photoshop where the backgrounds were deleted to create the cleanest image possible. Images were saved as Jpegs with 300 DPI. They are also available at 600 DPI. Images of five objects are only available at a lower resolution (200-500 pixels/inch) as these objects were no longer available for photographing, e.g., one object was a ceramic bookend that shattered shortly after the original photograph was taken. These objects have remained in the database for continuity as they were present in the original database. Each item in the database is assigned a unique, 4-digit ID number (cf. catalogue number) to facilitate communication between researchers. The sequence of the ID numbers was completely random.

Procedure and Design. Participants completed the experiment on lab computers either individually or in pairs (on separate machines on different sides of the lab). Before each task, written instructions were displayed on the screen. Participants first completed the novelty questions. Objects were displayed individually in the center of the screen in a random order. Below each object the question “Have you seen one of these before? (enter y/n)” was written. Participants responded by pressing the Y or N key. After each response the next object was displayed until participants responded to each of the 64 novel objects.

Next, participants answered the question “What would you call this object?” for each object. This question was asked to determine the degree of consensus among participants as what to call an object; we refer to this item as the “name question.” Again, objects were displayed individually in the center of the screen in a random order. Below each object there was a black text box in which participants could freely type their responses, which appeared in white font. Participants were instructed to type “XXX” if they wanted to skip an object (although the computer also accepted blank responses).¹ Participants pressed the return key to advance to the next object. After each response the next object was displayed until participants responded to each of the 64 novel objects.

Finally, participants answered the question “What do you really think this is?” for each object. This question was asked to determine the consensus regarding what each object really was; we refer to this item as the “identity question.” Objects and the free-response text box were displayed as in the previous task. Participants were instructed to type “XXX” if they did not have an answer (although the computer again accepted blank responses).² Participants pressed the return key to advance to the next object. After each response the next object was displayed until participants responded to each of the 64 novel objects. Objects were presented in a random order for each of the tasks.

Data were collected on a Dell computer. Each display was a 17-in. (43.18-cm) LCD monitor, with resolution set to 1280 x 1024 and refresh rate of 60 Hz. E-Prime (Version 2.0, Service Pack 1; Schneider, Eschman, & Zuccolotto, 2002) was used to control stimulus presentation and collect responses.

Coding. Verbatim responses to the name and identity questions were corrected for typos (e.g., letter omission: “ornge” to “orange,” letter order: “bule stuff” to “blue stuff,” wrong key on

keyboard: “squeexy toy” to “squeezy toy”) and spelling errors (e.g., “marakah” to “maraca,” “raidiator plug” to “radiator plug,” “aunement” to “ornament”—in this case the participant typed “aunement: sorry bad spelling”). In most cases corrections were facilitated because other participants provided correctly spelled responses (e.g., ornament) for the same item.

Circumlocution responses were coded as if the participant had used the noun he/she was describing, examples include: “help put on stubborn shoes” coded as “shoehorn,” “a toy to help children learn shapes: have to put 3D objects through the holes” coded as “shape sorter” and “a fluorescent multi-coloured object, which you could find in a fish tank, mainly fluorescent pink, also is shaped like the Sydney Opera House” coded as “fish tank accessory (fluorescent pink and multi-coloured)” (here “fish tank accessory” was used over “Sydney Opera House” because it occurred first in the response).

In many cases participants qualified their responses. However, because some participants spontaneously qualified their statements and others did not, qualifiers were not included when determining consensus (percent agreement). For example, for item 2033 “box,” “orange box,” “diamond box,” “orange diamond box” and “orange crate box” were all coded as agreeing on “box.” In the current study, consensus among the participants was more conservatively biased against our hypothesis that these are generally novel objects, therefore we coded these responses as if they were not qualified when calculating percent agreement. For example, both “alien looking thing” and “alien” were coded as “alien” (which increases participant consensus). This is in line with Landau, Smith and Jones (1988), who explain that object naming may be qualified, but the important content is the noun (which explains, among other things, why people refer to Philadelphia’s 60-foot sculpture of a clothespin as a “clothespin” although it cannot possibly be used for that function).

Finally for the consensus analyses, synonyms were collapsed, which increases participant agreement and is therefore again conservative against our hypothesis that it is difficult to agree on names for these items. Examples include: cylinder and tube; maraca and shaker; racket and paddle. This influenced name consensus for 39 objects ($M_{raw} = 35\%$, $SD_{raw} = 14\%$; $M_{adjusted} = 47\%$, $SD_{adjusted} = 19\%$) and identity consensus for 33 objects ($M_{raw} = 38\%$, $SD_{raw} = 17\%$; $M_{adjusted} = 49\%$, $SD_{adjusted} = 20\%$). Consensus scores were calculated out of the number of participants who provided a response for a given object. For example, 11 participants responded that object 2010 really is a bone ($n = 1$) or dog toy ($n = 10$), but only 21 participants provided a response for that object, yielding a 52% agreement. Again, this is a more conservative approach than taking the absolute agreement (e.g., 11 out of 32 is 34%).

Results and Discussion

The objects in the NOUN Database are generally novel as indicated by self-report: $M_{novelty} = 69\%$, $SD_{novelty} = 19\%$, range = 19%-97%. This novelty is also reflected in the (lack of) consensus on what to call the objects when they should be passed ($M = 47\%$, $SD = 19\%$) and the (lack of) consensus on what the objects really are ($M = 49\%$, $SD = 20\%$). These rates are significantly less than the 85% agreement threshold set by Samuelson and Smith (1999), both $t_s > 14.20$ and $p_s < .001$, two-tailed, both $d_s > 1.76$. Using Samuelson and Smith's (1999) threshold, participants agreed on what to call objects 2002, 2003, 2032, 2059, but only when we collapsed across synonyms. Note, we were particularly lenient in accepting synonyms for object 2032 and included any names for vehicles or flying objects. Using the same threshold, participants agreed on what objects 2032 and 2059 really were, but again only when we collapsed across synonyms.

Novelty scores were negatively correlated with both name consensus ($r = -.290, p = .02$, 95% CI = $-.048$ to $-.500$) and identity consensus ($r = -.465, p < .001$, 95% CI = $-.247$ to $-.638$). That is, the more novel the object the less likely participants agreed on the name or identity for the object. If an object is familiar it should be easier to have agreement on what it is and what to call it (especially when collapsed across synonyms, which we have done). Thus, these negative correlations provide additional evidence that the novelty scores are reliable.

When viewing the free responses during the consensus analyses, the use of qualifiers was staggering. We would be remiss if we did not disseminate these findings. In total, participants spontaneously provided 1,426 color and texture qualifiers in their statements, although each object was presented separately on a decontextualized background. For each object, the proportion of colors and proportion of textures (e.g., spikey, soft) for the name and identity questions were calculated as the number of qualifiers given the number of responses. Proportions of qualifiers were submitted to a qualifier type (color, texture) by trial type (name question, identity question) repeated-measures ANOVA. The ANOVA yielded a significant qualifier type by trial type interaction, $F(1,63) = 159.19, p < .001, \eta_p^2 = .72$ (see Figure 1), indicating that participants were both significantly more likely to use color terms ($M = .27, SD = .15$) than texture terms ($M = .10, SD = .14$) and significantly more likely to use qualifiers when asked what they would call something ($M = .33, SD = .15$) than when they were asked what something really is ($M = .04, SD = .07$). Significant main effects for qualifier type, $F(1,63) = 88.30, p < .001, \eta_p^2 = .58$, and trial type, $F(1,63) = 617.72, p < .001, \eta_p^2 = .91$ were also found. The greater use of qualifiers on the name trials may reflect participants' uncertainty when asked what to name something novel. Indeed, proportion of color qualifiers on the naming trials was significantly correlated with novelty, $r = .42, p = .0006, CI = .189 - .599$. Thus, when participants do not

know what something is they rely on color to refer to the object. This pattern of responding can also be seen if we compare the number of color qualifiers for self-reported novel versus known objects, $t(63) = 7.93$, $p < .0001$, two-tailed, $d = 1.14$.

Experiment 2

In Experiment 2, our participants provided similarity ratings on the NOUN stimuli, which were subjected to a multidimensional scaling (MDS) analysis. It should be noted that MDS is not the only approach we could have adopted. There are a variety of models of similarity and each has its own assumptions regarding the fundamental ways in which psychological similarity is represented by people, the ways in which similarity estimates are constructed, and so on (see Hahn, 2014, for discussion). Spatial models of similarity represent objects as points in a hypothetical “psychological space,” wherein the similarity of a pair of items is represented by their distance in space (with like items being located close to one another, and vice versa; see Shepard, 1980; Shepard, 1987). MDS techniques are specifically designed to “uncover” these psychological maps, and have even been successfully incorporated into sophisticated mathematical models of cognition, such as the *Generalized Context Model* (GCM; Nosofsky, 1986).

An alternative approach to adopting a spatial model of similarity would be to examine similarity from the perspective of featural accounts, which assume that the basic representational units of similarity are not continuous, but are binary. Proponents of featural accounts point out that spatial models have some problematic assumptions, such as symmetry constraints (e.g., a Chihuahua may be rated as more similar to a Labrador than a Labrador is to a Chihuahua) and the triangle inequality axiom (see, Goldstone & Son, 2012; Tversky, 1977). Statistical techniques

like *additive clustering* (e.g., Shepard & Arabie, 1979) are typically adopted when the analyst assumes a featural account of similarity. More recent approaches have even extended these tools to accommodate continuous dimensions and discrete features using Bayesian statistics (see, Navarro & Griffiths, 2008; Navarro & Lee, 2003, 2004).

Our primary reasons for adopting a spatial model of similarity relate to the novelty and real-world nature of the NOUN stimuli. With featural models, detailed predictions about similarity structure are sometimes difficult to make because specific assumptions need to be made regarding the features that comprise the objects. The NOUN stimuli are complex, real-world objects, but their novelty makes it difficult to make any predictions regarding what features from which they may be comprised. Spatial models (implemented using MDS), however, are agnostic about these features. Moreover, upon inspection of the objects themselves, it seems more likely that the salient features are continuous, rather than binary, making a spatial model more attractive to adopt. For instance, because many of the objects are brightly multi-colored, a continuous color dimension seems more appropriate than a featural representation (e.g., an object may be entirely green, mostly green, partially green, and so on). Finally, we find spatial models appealing in general, because of their successful incorporation into exemplar models of cognition. Exemplar models, such as the GCM (Nosofsky, 1986), are adept at capturing human categorization and recognition behavior, and these are precisely the types of issues to which we hope future researchers will utilize the NOUN stimuli.

Method

Participants. The same participants as in Experiment 1 took part in Experiment 2. This experiment took place immediately after Experiment 1.

Materials. The same stimuli as in Experiment 1 were used.

Procedure and Design. Participants were shown all 64 objects across 13 trials, and provided similarity ratings using the *spatial arrangement method* (*SpAM*, Hout, Goldinger, & Ferguson, 2013, see also Goldstone, 1994). On each trial, twenty different pictures were shown to the participant, arranged in 4 discrete rows (with 5 items per row), with random item placement. Participants were instructed to drag and drop the images in order to organize the space such that the distance among items was proportional to each pair's similarity (with items closer in space denoting greater similarity). Participants were given as much time as they needed to arrange the space on each trial; typically, trials lasted between 2 and 5 minutes. Once participants finished arranging the items, they completed the trial by clicking the right mouse-button. To avoid accidental termination, participants were asked if the space was satisfactory (indicating responses via the keyboard) and were allowed to go back and continue the trial if desired.

The x- and y-coordinates of each image was then recorded and the Euclidean distance (in pixels) between each pair of stimuli was calculated (for 20 stimuli there are 190 pairwise Euclidean distances). This procedure was performed repeatedly (over 13 trials), but with different images presented together on each trial, so that all pairwise comparisons among the 64 total images were recorded. Thus, this provided a full similarity matrix comparing the ratings of each image to all of the other images (i.e., all 2016 comparisons) for each participant. This took participants under an hour to complete; similar rating procedures have been used by other researchers (Goldstone, 1994; Hout, Papesh, & Goldinger, 2012; Kriegeskorte & Marieke, 2012).

Stimulus selection. We controlled the selection of images on each trial by employing a Steiner System (see Doyen, Hubaut, & Vandensavel, 1978); these are mathematical tools that can be used to ensure that each item in a pool is paired with every other item (across subsets/trials) at least once. A Steiner System is often denoted $S(v, k, t)$, where v is the total number of stimuli, k is the number of items in each subset, and t is the number of items that need to occur together. Thus for us, v , k , and t , are 64 (total images), 20 (images per trial), and 2 (denoting pairwise comparisons), respectively. Simply put, the Steiner System provides a list of subsets (i.e., trials) identifying which items should be presented together on each trial. For some combinations of v and k , there may exist a Steiner set that does not repeat pairwise comparisons (i.e., each pair of items is shown together once and only once). For other combinations (including ours), some stimuli must be shown with others more than once. Because this leads to multiple observations per “cell,” we simply took the average of the ratings for the pairs that were shown together more than once. Across participants, images were randomly assigned to numerical identifiers in the Steiner System, which ensured that each participant saw each pair of images together at least once, but that different people received different redundant pairings (see also Berman et al., 2014, for similar use of multi-trial implementations of the spatial arrangement method).

MDS Analysis. After the similarity matrices were compiled, we performed multidimensional scaling on the pairwise Euclidean distances, using the PROXSCAL scaling algorithm (Busing, Commandeur, Heiser, Bandilla, & Faulbaum, 1997). Data were entered into SPSS in the form of individual similarity matrices for each participant; averaging was performed by the PROXSCAL algorithm automatically. To determine the correct dimensionality of the space, we created Scree plots, plotting the model’s stress against the number of dimensions used

in the space (see also Hout, Papesh, & Goldinger, 2012). Stress functions vary across scaling algorithms (PROXSCAL uses “normalized raw stress”), but all are computed to measure the agreement between the estimated distances provided by the MDS output and the raw input proximities (lower stress values indicate a better model fit). Scree plots are often used to determine the ideal dimensionality of the data by identifying the point at which added dimensions fail to improve the model fit substantially (Jaworska & Chupetlovska-Anastasova, 2009).

PROXSCAL offers several options for determining the starting configuration of the MDS space (prior to the iterative process of moving the points around in space to improve model fit). Specifically, there are Simplex, Torgerson, and Multiple Random Starts options. The first two options use mathematical principles to determine the starting configuration, whereas the latter takes multiple attempts at scaling, using completely random configurations (the output with the lowest stress value is then reported). The benefit of the Multiple Random Starts approach is that with deterministic starting configurations (like those used by Simplex and Torgerson algorithms), as the iterative process of moving the items in the space progresses, there is a risk that the solution will fall into a local (rather than a global) minimum with respect to stress. Because the process is repeated many times under a Multiple Random Starts approach, that risk is reduced. We created Scree plots using each of the options (for Multiple Random Starts we implemented 100 random starting configurations per dimensionality).

Results and Discussion

Scree Plots. Scree plots for each of the starting configuration options are shown in Figure 2. It is clear from these plots that the data are comprised of more than just two primary featural

dimensions. Each space exhibits a remarkably similar shape, with stress values beginning to plateau at around four dimensions.

Configuration. In order to choose the “best” solution to report, we correlated the inter-item distance vectors across solutions derived from each of the three starting configurations (see Hout, Goldinger, and Ferguson, 2012 and Goldstone, 1994 for similar approaches). This provides a metric to gauge the extent to which the arrangement of points in one MDS space “agrees” with the others (e.g., a pair of points that is located close together in one space should be close together in the others, and vice versa). Seeing that stress values were nearly identical across solutions (see Figure 2), we then chose to report the configuration derived from the Torgerson starting configuration because its organization correlated the most strongly with the other two (Pearson correlation coefficients with Simplex and Multiple Random Starts configurations were .72 and .75, respectively, indicating a high degree of agreement across spaces). Below, we provide the scaling solution based on the Torgerson option in four-dimensional space. The results are shown in Figures 3 and 4. In the figures, the objects are superimposed on the resulting MDS plot such that they are placed based on their weights on dimensions 1 and 2 (Figure 3) or dimensions 3 and 4 (Figure 4).

Item pairings, sorted by similarity. With the coordinates obtained from the MDS space, it is possible to identify object pairs that are more or less similar to one another, relative to the other possible pairs. No basic unit of measurement is present in MDS, so the inter-item distances are expressed in arbitrary units. This means it is not possible to define numerical cutoff values for classifying pairs as “very similar,” “very dissimilar,” and so on. In order to provide empirically-driven identification of item pairs, we first created a vector of distances, corresponding to the 2016 Euclidean distances (in four-dimensional psychological space) for all

item pairs. Next, the distances were rank-ordered, and categorized based on a quartile-split. The 504 image pairs with the closest MDS distances to one another were classified in the first quartile. The next 504 rank-ordered image pairs were classified in the second quartile, and so on. For each pair of images, we provide the Euclidean distance in four-dimensional space, the ordinal ranking of the pair (where 1 is the most similar pair, and 2016 is the most distal pair), and the classification of the pair (first, second, third, or fourth quartile). This classification is provided as a convenience to researchers who wish to easily identify item pairs that vary along a continuum of similarity (see Hout et al., 2014, for a similar approach). When this classification system was applied to our stimuli, we found mean MDS distances of 0.5181 for the first quartile ($SD = 0.1128$, range = 0.1654 - 0.6871), 0.7973 for the second quartile ($SD = 0.0598$, range = 0.6876 - 0.8877), 0.9711 for the third quartile ($SD = .0482$, range = 0.8879 - 1.0533), and 1.1400 for the fourth quartile ($SD = 0.0641$, range = 1.0535 - 1.3593).

Figure 5 displays item pairings, sampled from each of the four quartiles. Shown in the top-left and bottom-right of the figure are the most and least similar items pairs, respectively. The bottom-left and top-right of the figure shows pairings sampled from the top and bottom of the second and third quartiles, respectively. It is clear that the items become less similar across quartiles. The first quartile pair exhibit a strikingly similar shape and texture. The second and third quartile pairings are less alike, with similarities in overall shape for the second quartile pair, and only a vague sense of sameness for the third quartile. The fourth-quartile pair are clearly the least similar, with no discernable likenesses in shape, texture, color, etc.

Finally, as an aid to other researchers, we determined the 16 items that were the most and the least similar to the set as a whole. Specifically, for each object we calculated the mean of the MDS distances between that object and each of the other 63 objects. The overall grand mean was

.8566 ($SD = .0367$, range = 0.7546 - 0.9348). The 16 items with the smallest mean distances (i.e., the items most similar to the sample as a whole) and the 16 items with the largest mean distances (i.e., the items most dissimilar to the collection as a whole) are presented in Table 1.

Insert Table 1 About Here

Experiment 3

In the categorization literature, there is often higher within-category similarity for basic-level categories than for global-level categories (see e.g., Rosch, 1978 for a discussion). Consider for example the nested categories of birds, owls and snowy owls. All birds share several features, in that they breathe, eat food, have feathers, etc. All owls also share features (e.g., they are predatory, they fly). Finally, all snowy owls share very specific features (e.g., yellow eyes, black beaks). At the lowest level, category members are most likely to be considered similar to one another. Moving up the hierarchy, categories become more inclusive (barn owls and great horned owls are also owls; penguins and hummingbirds are also birds) and category members become less similar as they share fewer features. Thus, our main goal with Experiment 3 was to determine whether the categories in the NOUN Database reflect this relationship between categories at different levels. In particular, we use between-category distances to determine which items, if any, belong to the same global-level categories.

Method

Participants. The same participants that took part in Experiments 1 and 2 also completed Experiment 3. This experiment took place immediately after Experiment 2.

Stimuli. Participants saw three exemplars for ten of the object categories seen in Experiments 1 and 2. We selected exemplars that only differed from each other in color, as this

is a common method for forming categories for experimental research (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Twomey et al., 2014; Vlach, Sandhofer, & Kornell, 2008; Woodward, Markman, & Fitzsimmons, 1994). These images were created in identical fashion to Experiment 1.

Procedure and Design. The procedure was identical to Experiment 2 except that all 30 objects were displayed on the computer screen for a single SpAM trial. We included all 30 items at once so that none of the subcategories was more familiar or novel than any other category when the similarity judgments were made, and so that the overall context of the similarity judgments was the same for all participants.

Results and Discussion

In order to screen our data for potential outliers, we analyzed the extent to which individual participants' MDS spaces correlated with all others. Because similarity is a subjective construct, it is to be expected that individual participant's solutions will deviate from other people. The following approach is simply a coarse measure, designed to identify participants who may not have been taking the task seriously (and therefore are only contributing noise to the data). Conceptually, this is akin to studies of reaction time (RT) wherein participants with exceedingly long average RTs are removed to provide a cleaner estimate of the true response time.

Our approach entailed several steps (see Hout et al., 2013, for a similar approach); it should be noted that we applied this same criterion to the data in Experiment 2, but no participants were deemed outliers in that experiment. First, we created individual MDS spaces for each participant and derived vectors of inter-item distances from those spaces. Second, we correlated the distance vectors across all participants. Finally, we calculated the average

correlation coefficient for each participant. We excluded one participant from further analysis for having an average correlation coefficient that was more than 2.5 standard deviations below the mean of the group.

Scree Plots. Scree plots for each of the starting configuration options are shown in Figure 6. As before, the data are clearly comprised of more than two featural dimensions, with stress values flattening out at around four dimensions.

Configuration. As before, we correlated the inter-item distance vectors across solutions derived from each of the three possible starting configurations to choose the “best” configuration. Stress values were again comparable across configurations, and again, the Torgerson solution correlated most strongly with the other two (Pearson correlation coefficients with Simplex and Multiple Random Starts configurations were .93 and .89, respectively, indicating once more a high degree of agreement across spaces). Thus, we provide the scaling solution based on the Torgerson option in four-dimensional space, with results shown in Figures 7 and 8. In the figures, the objects are superimposed on the resulting MDS plot such that they are placed based on their weights on dimensions 1 and 2 (Figure 7) or dimensions 3 and 4 (Figure 8).

Global category classification. We next compared the mean distances between the different subcategories (see Suppl. Table 1), both within and between categories. Distances ranged from 0.06 to 1.21 in multidimensional space ($M_{distance} = 0.78$, $SD_{distance} = 0.37$). Recall, there is no basic unit of measurement in MDS. Within-category distances were low ($M_{distance} = 0.11$, $SD_{distance} = 0.04$, range = .06-.19); that is, items were considered highly similar, which we should expect because the subcategories consisted of exemplars that only varied in their color makeup. Recall, our main goal was to use between-category distances to determine if any items were considered to belong to the same global-level categories. Between-category distances were

higher than within-category distances ($M_{distance} = 0.93$, $SD_{distance} = 0.20$, range = 0.39-1.21). We sought a relatively conservative cut-off for classifying subcategories as belonging to the same global-level category and chose $M_{distance} + .25SD_{distance}$ (0.87). Using this criterion, the 10 subcategories formed nine global categories (Figure 9). Notice that two categories included 3 basic-level categories.

Before we continue, we would like to discuss the “messiness” among the global-level categories. For example, subcategories 2015 and 2039 form a global-level (superordinate) category and 2051 and 2039 form a global-level category, but 2015 and 2051 do not form a category. We maintain that such oddities reflect the nature of global-level categories. Consider for example airbus aircrafts, snowy owls and cruise ships. One might consider airbus aircrafts and snowy owls to be members of the same global-level category, because they both fly, and airbus aircrafts and cruise ships as members of the same global-level category, because they are both large passenger transport vehicles. But one would likely not consider snowy owls and cruise ships to form a global-level category. The key point is that one’s notion of similarity changes as a function of the context in which judgments are being made, and thus the relationships among object categories sometimes overlap. Thus, the “messiness” in the global-level category structure is not problematic, per se. Rather it is consistent with the categorization and similarity literatures (see, Goldstone, Medin, & Gentner, 1991; Hout et al., 2013). Therefore, we sought to confirm that the categories in the NOUN Database reflected this relationship between categories at different levels. To this end we submitted the distances to a one-way ANOVA with three levels of categories: basic, global and unrelated (e.g., 2039 and 2040). The ANOVA yielded a main effect of category, $F(2,54) = 326.10$, $p < .001$, $\eta_p^2 = .93$. Follow-up Tukey’s HSD revealed that there was greater within-category similarity for the basic-level exemplars ($M_{basic} = 0.1150$,

$SD_{basic} = 0.0408$, range = .0610 = .1894) than for the global-level exemplars ($M_{global} = 0.7063$, $SD_{global} = 0.1489$, range = 0.3914 = 0.8708, $p < .01$) and than for the unrelated exemplars ($M_{unrelated} = 1.0517$, $SD_{unrelated} = 0.0908$, range = 0.8911 = 1.2144, $p < .01$). The global-level exemplars also had higher within-category similarity than the unrelated exemplars ($p < .01$). Recall that smaller numbers indicate participants placed the items in closer spatial proximity (i.e., smaller numbers indicate greater similarity). Thus, the structure of these categories is similar to other categories used in experimental research.

General Discussion

In this contribution, we have effectively expanded the Novel Object & Unusual Name (NOUN) Database. The current series of experiments confirm that the objects included are in fact novel and can be grouped into basic- and global-level categories. Additional analyses revealed a negative relationship between novelty and consensus on what to call the objects and what they really are. Further, the more novel the objects were, the more likely participants were to qualify their responses, particularly by mentioning the colors of the objects.

The NOUN Database presents a ready-to-use collection of images of real 3D objects. This collection of naturalistic stimuli works well alone or against images of known objects as these objects have realistic shading, colors, textures and complexity—often including multiple parts. The novelty ratings obtained in Experiment 1 and the similarity ratings obtained in Experiments 2 and 3 can inform researchers' decisions on which stimuli to use depending on their research questions. For example, whether they require stimulus sets with minimal differences in novelty across items or with maximal dissimilarity between items. Thus, the use of a pre-existing database, like the NOUN Database, may help researchers to save time and cut

research expenses. Importantly, the similarity ratings of multiple exemplars from ten categories (Experiment 3) provides useful information for researchers who wish to conduct experiments involving category learning and want to avoid simply relying on their own intuitive judgments of the relationships between various novel object categories.

Novel objects are useful for a variety of research designs (e.g., Axelsson & Horst, 2013, 2014; Kwon et al., 2014) and multiple exemplars of novel objects may be useful for categorization (e.g., Homa et al., 2011; Twomey et al., 2014), memory (Hout & Goldinger, 2010, 2012) and perception (e.g., Cunningham & Wolfe, 2014; Hout & Goldinger, 2015) experiments. Importantly, the objects in the NOUN Database were generally novel and not already associated with agreed-upon names. As such, these objects are particularly well-suited for studies in which participants are learning name-object associations for the first time (e.g., Horst et al., 2011). Specifically, the use of such novel stimuli can help ensure observed effects are due to the experimental manipulation and not a priori knowledge (see also, Ard & Beverly, 2004; Bornstein & Mash, 2010). In the real world, particularly for children, all objects were once novel, unnamed objects (Horst, 2013; Horst, McMurray, & Samuelson, 2006)—only through cumulative, gradual learning do we learn to attribute identities to the objects around us. Therefore, using novel objects and pseudo-word names is an ecologically valid design principle in a variety of settings.

The NOUN Database is a resource, which will enable psychologists and empiricists in other domains to obtain the best possible experimental control in situations wherein real-world stimuli are required, but prior knowledge regarding the identity of the objects is potentially problematic. As such, this tool should allow researchers to increase the confidence in their experimental findings, and we hope it will help advance our collective scientific understanding of the processes underlining word learning, memory consolidation, categorization and more.

References

- Ard, L. M., & Beverly, B. L. (2004). Preschool word learning during joint book reading: Effect of adult questions and comments. *Communication Disorders Quarterly*, 26(1), 17-28.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2), 108-119. doi: 10.1002/per.1919
- Axelsson, E. L., & Horst, J. S. (2013). Testing a word is not a test of word learning. *Acta Psychologica*, 144(2), 264-268. doi: 10.1016/j.actpsy.2013.07.002
- Axelsson, E. L., & Horst, J. S. (2014). Contextual Repetition Facilitates Word Learning via Fast Mapping. *Acta Psychologica*, 152, 95-99. doi: 10.1016/j.actpsy.2014.08.002
- Berman, M. G., Hout, M. C., Kardan, O., Hunter, M., Yourganov, G., Henderson, J. M., . . . Jonides, J. (2014). The perception of naturalness converges with objective measures of naturalness. *PloS One*, 9(e114572). doi: 10.1371/journal.pone.0114572.
- Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Referent selection and word learning in 18- and 24-month-old infants. *Cognition*, 126(1), 39-53. doi: 10.1016/j.cognition.2012.08.008
- Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, 81(3), 881-897.
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 105(38), 14325-14329. doi: 10.1073/pnas.0803390105
- Busing, F. M. R. A., Commandeur, J. J. F., Heiser, W. J., Bandilla, W., & Faulbaum, F. (1997). PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla & F. Faulbaum (Eds.), *Advances in Statistical Software* (Vol. 6, pp. 67-73). Stuttgart: Lucius & Lucius.
- Cunningham, C. A., & Wolfe, J. M. (2014). The role of object categories in hybrid visual and memory search. *Journal of Experimental Psychology: General*, 143, 1585-1599. doi: 10.1037/a0036313
- Dan-Glauser, E. S., & Scherer, K. R. V. (2011). The Geneva Affective Picture Database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2), 468-477. doi: 10.3758/s13428-011-0064-1
- Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1), 351-362. doi: 10.3758/BRM.42.1.351
- Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12), 1673-1682. doi: 10.1016/S0042-6989(96)00286-6
- Gentner, D. (1978). A study of early word meaning using artificial objects: What looks like a jiggy but acts like a zimbo. *Papers and Reports on Child Language Development*, 15, 1-6.
- Giezen, M. R., Escudero, P., & Baker, A. E. (in press). Rapid learning of minimally different words in five- to six-year-old children: Effects of acoustic salience and hearing impairment. *Journal of Child Language*. doi: 10.1017/S0305000915000197
- Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4), 381-386. doi: 10.3758/BF03204653

- Goldstone, R., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, 23(2), 222-262. doi: 10.1016/0010-0285(91)90010-L
- Goldstone, R., & Son, J. Y. (2012). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 155-176). New York: Oxford University Press.
- Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young-Children and Adults Use Lexical Principles to Learn New Nouns. *Developmental Psychology*, 28(1), 99-108. doi: 10.1037/0012-1649.28.1.99
- Hahn, U. (2014). Similarity. *WIREs Cognitive Science*, 5(3), 271-280. doi: 10.1002/wcs.1282
- Homa, D., Hout, M. C., Milliken, L., & Milliken, A. M. (2011). Bogus concerns about the false prototype enhancement effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2), 368-377. doi: 10.1037/a0021803
- Horst, J. S. (2009). Novel Object & Unusual Name (NOUN) Database [PDF document]. Retrieved 6/6/14 from <http://www.sussex.ac.uk/wordlab/noun>.
- Horst, J. S. (2013). Context and Repetition in Word Learning. *Frontiers in Psychology*, 4(149), 1-11. doi: 10.3389/fpsyg.2013.00149
- Horst, J. S., McMurray, B., & Samuelson, L. K. (2006). Online Processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 339-344). Austin: Tx: Cognitive Science Society.
- Horst, J. S., & Samuelson, L. K. (2008). Fast Mapping But Poor Retention By 24-Month-Old Infants. *Infancy*, 13(2), 128-157. doi: 10.1080/15250000701795598
- Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition*, 118(2), 234-244. doi: 10.1016/j.cognition.2010.10.015
- Horst, J. S., & Twomey, K. E. (2013). It's taking shape: Shared object features influence novel noun generalizations. *Infant and Child Development*, 22(1), 24-43. doi: 10.1002/icd.1768
- Hout, M. C., & Goldinger, S. D. (2010). Learning in repeated visual search. *Attention, Perception & Psychophysics*, 72(5), 1267-1282. doi: 10.3758/APP.72.5.1267
- Hout, M. C., & Goldinger, S. D. (2012). Incidental learning speeds visual search by lowering response thresholds, not by improving efficiency: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1), 90-112. doi: 10.1037/a0023894
- Hout, M. C., & Goldinger, S. D. (2015). Target templates: The Precision of Mental Representations Affects Attentional Guidance and Decision-Making in Visual Search. *Attention, Perception & Psychophysics*, 2015(1), 129-149. doi: 10.3758/s13414-014-0764-6
- Hout, M. C., Goldinger, S. D., & Brady, K. J. (2014). MM-MDS: A multidimensional scaling database with similarity ratings for 240 object categories from the Massive Memory picture database. *PloS One*, 9(11), e112644. doi: 10.1371/journal.pone.0112644
- Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1), 256-281. doi: 10.1037/a0028860

- Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2012). Multidimensional scaling. *Wiley Interdisciplinary Reviews (WIREs): Cognitive Science*, 4(1), 93-103. doi: 10.1002/wcs.1203
- Imai, M., & Gentner, D. (1997). A Cross-Linguistic Study of Early Word Meaning: Universal Ontology and Linguistic Influence. *Cognition*, 62(2), 169-200. doi: 10.1016/S0010-0277(96)00784-6
- Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods of Psychology*, 5, 1-10.
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558-578. doi: 10.1037/a0019165
- Kucker, S. C., & Samuelson, L. K. (2012). The first slow step: Differential effects of object and word- form familiarization on retention of fast- mapped words. *Infancy*, 17(3), 295-323. doi: 10.1111/j.1532-7078.2011.00081.x
- Kwon, M.-K., Luck, S. J., & Oakes, L. M. (2014). Visual short-term memory for complex objects in 6- and 8-month-old infants. *Child Development*, 85(2), 564-577. doi: 10.1111/cdev.12161
- Landau, B., Smith, L. B., & Jones, S. (1998). Object shape, object function, and object name. *Journal of Memory & Language*, 38(1), 1-27.
- Lipinski, J., & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory and Language*, 52(2), 171-192. doi: 10.1016/j.jml.2004.10.004
- Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, 46(2), 596-610. doi: 10.3758/s13428-013-0379-1
- Mather, E., & Plunkett, K. (2009). Learning Words Over Time: The Role of Stimulus Repetition in Mutual Exclusivity. *Infancy*, 14(1), 60-76.
- Mather, E., Schafer, G., & Houston-Price, C. (2011). The impact of novel labels on visual processing during infancy. *British Journal of Developmental Psychology*, 29(4), 783-805. doi: 10.1348/2044-835X.002008
- Matheson, H. E., & McMullen, P. A. (2011). A computer-generated face database with ratings on realism, masculinity, race, and stereotypy. *Behavior Research Methods*, 43(1), 224-228. doi: 10.3758/s13428-010-0029-9
- Migo, E. M., Montaldi, D., & Mayes, A. R. (2013). A visual object stimulus database with standardized similarity information. *Behavior Research Methods*, 45(2), 344-354. doi: 10.3758/s13428-012-0255-4
- Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, 20(1), 2597-2628. doi: 10.1162/neco.2008.04-07-504
- Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems* (pp. 59-66). Cambridge, MA: MIT Press.

- Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961-974. doi: 10.3758/BF03196728
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57. doi: 10.1037/0096-3445.115.1.39
- Rakison, D. H., & Poulin-Dubois, D. (2002). You go this way and I'll go that way: Developmental changes in infants' detection of correlations among static and dynamic features in motion events. *Child Development*, *73*(3), 682-699. doi: 10.1111/1467-8624.00432
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Llyod (Eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339-349.
- Samuelson, L. K., & Horst, J. S. (2007). Dynamic Noun Generalization: Moment-to-Moment Interactions Shape Children's Naming Biases. *Infancy*, *11*(1), 97-110. doi: 10.1207/s15327078in1101_5
- Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1-33. doi: 10.1016/S0010-0277(99)00034-7
- Samuelson, L. K., & Smith, L. B. (2005). They Call It like They See It: Spontaneous Naming and Attention to Shape. *Developmental Science*, *8*(2), 182-198. doi: 10.1111/j.1467-7687.2005.00405.x
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime computer software and manual*. Pittsburgh, PA: Psychology Software Tools, Inc.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398. doi: 10.1126/science.210.4468.390
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323. doi: 10.1126/science.3629243
- Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123. doi: 10.1037/0033-295X.86.2.87
- Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources. *Journal of General Internal Medicine*, *26*(8), 920-929. doi: 10.1007/s11606-010-1621-5
- Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800-811. doi: 10.1037/0278-7393.28.4.800
- Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568. doi: 10.1016/j.cognition.2007.06.010
- Soja, N. N., Carey, S., & Spelke, E. S. (1992). Perception, ontology, and word meaning. *Cognition*, *45*(1), 101-107. doi: 10.1016/0010-0277(92)90025-D
- Tversky, A. (1977). Features of similarity. *84*, 4(327-352). doi: 10.1037/0033-295X.84.4.327

- Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's More Like It: Multiple Exemplars Facilitate Word Learning. *Infant and Child Development*, 23(2), 105-122. doi: 10.1002/icd.1824
- Umla-Runge, K., Zimmer, H. D., Fu, X., & Wang, L. (2012). An action video clip database rated for familiarity in China and Germany. *Behavior Research Methods*, 44(4), 946-953. doi: 10.3758/s13428-012-0189-x
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words.
- View Profile; Luce, Paul A.View Profile. *Psychological Science*9.4 (Jul 1998): 325-329. . *Psychological Science*, 9(4), 325-329. doi: 10.1111/1467-9280.00064
- Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, 109, 163-167. doi: 10.1016/j.cognition.2008.07.013
- Warren, D. E., & Duff, M. C. (2014). Not so fast: Hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus*, 24(8), 920-933. doi: 10.1002/hipo.22279
- Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, 34(6), 1289-1309.
- Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, 30(4), 553-566. doi: 10.1037/0012-1649.30.4.553
- Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people10.1037/a0024023. *Developmental Psychology*, 47(5), 1220-1229. doi: 10.1037/a0024023
- Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, 18(5), 414-420. doi: 10.1111/j.1467-9280.2007.01915.x
- Zosh, J. M., Brinster, M., & Halberda, J. (2013). Optimal Contrast: Competition Between Two Referents Improves Word Learning. *Applied Developmental Science*, 17(1), 20-28. doi: 10.1080/10888691.2013.748420

Footnotes

1 One participant entered “XXX” 11 times. This participant and four others left answers blank a total of 54 times ($M_{non-responses} = 10.8$, $SD_{non-responses} = 11.9$, range 1-29 for these 5 participants). In total, for all 32 participants, only 2% of the description questions were unanswered.

2 In total 26 of the 32 participants entered “XXX” or a blank response for at least 1 item ($M_{non-responses} = 32.62$, $SD_{non-responses} = 19.04$, range 1-64 for these participants). In total, for all 32 participants, 21% of the identity questions were unanswered.

Supplemental Files

1. NOUN Database 2nd Edition:

<https://www.sussex.ac.uk/webteam/gateway/file.php?name=noun-database-2nded-a4.pdf&site=390>

2. Sorting Table: <http://www.sussex.ac.uk/wordlab/documents/noun-sorting-tables.xlsx>

Table 1. The MDS distances for the 16 most similar and 16 least similar objects, see the NOUN Database for images of these objects.

	Rank	Item	Mean Distance
Most similar items	1	2030	0.7546
	2	2022	0.7620
	3	2032	0.7721
	4	2019	0.7758
	5	2011	0.7956
	6	2007	0.7977
	7	2026	0.8104
	8	2039	0.8123
	9	2038	0.8131
	10	2008	0.8227
	11	2056	0.8295
	12	2043	0.8298
	13	2014	0.8307
	14	2042	0.8322
	15	2013	0.8344
	16	2016	0.8394
Most dissimilar items	49	2061	0.8781
	50	2062	0.8793
	51	2010	0.8798
	52	2040	0.8808
	53	2054	0.8811
	54	2037	0.8826
	55	2028	0.8871
	56	2063	0.8886
	57	2023	0.8923
	58	2020	0.8943
	59	2045	0.8984
	60	2002	0.9052
	61	2005	0.9095
	62	2031	0.9176
	63	2057	0.9307
	64	2046	0.9348

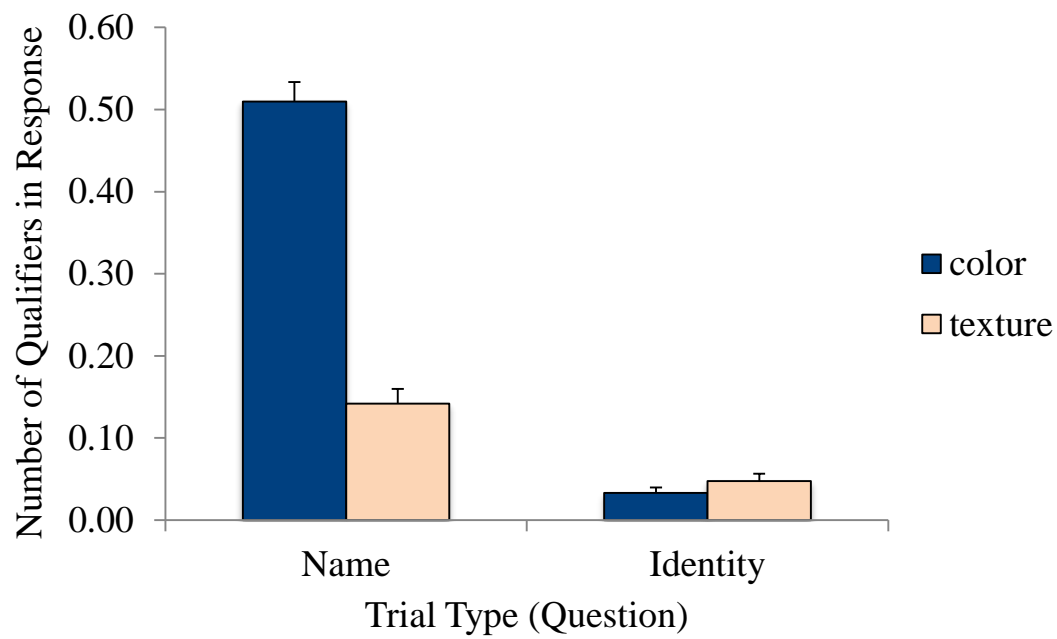


Figure 1. Number of qualifiers as a function of trial type (question), from Experiment 1. Error bars represent one standard error.

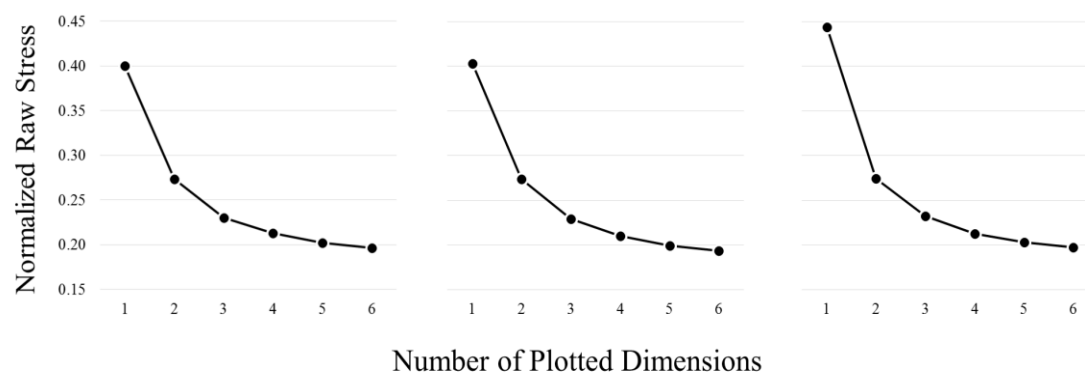


Figure 2. Scree plots for the Simplex (left), Torgerson (middle), and Multiple Random Starts (right) options, from Experiment 2. Stress values are plotted as a function of the dimensionality in which the MDS data were scaled.

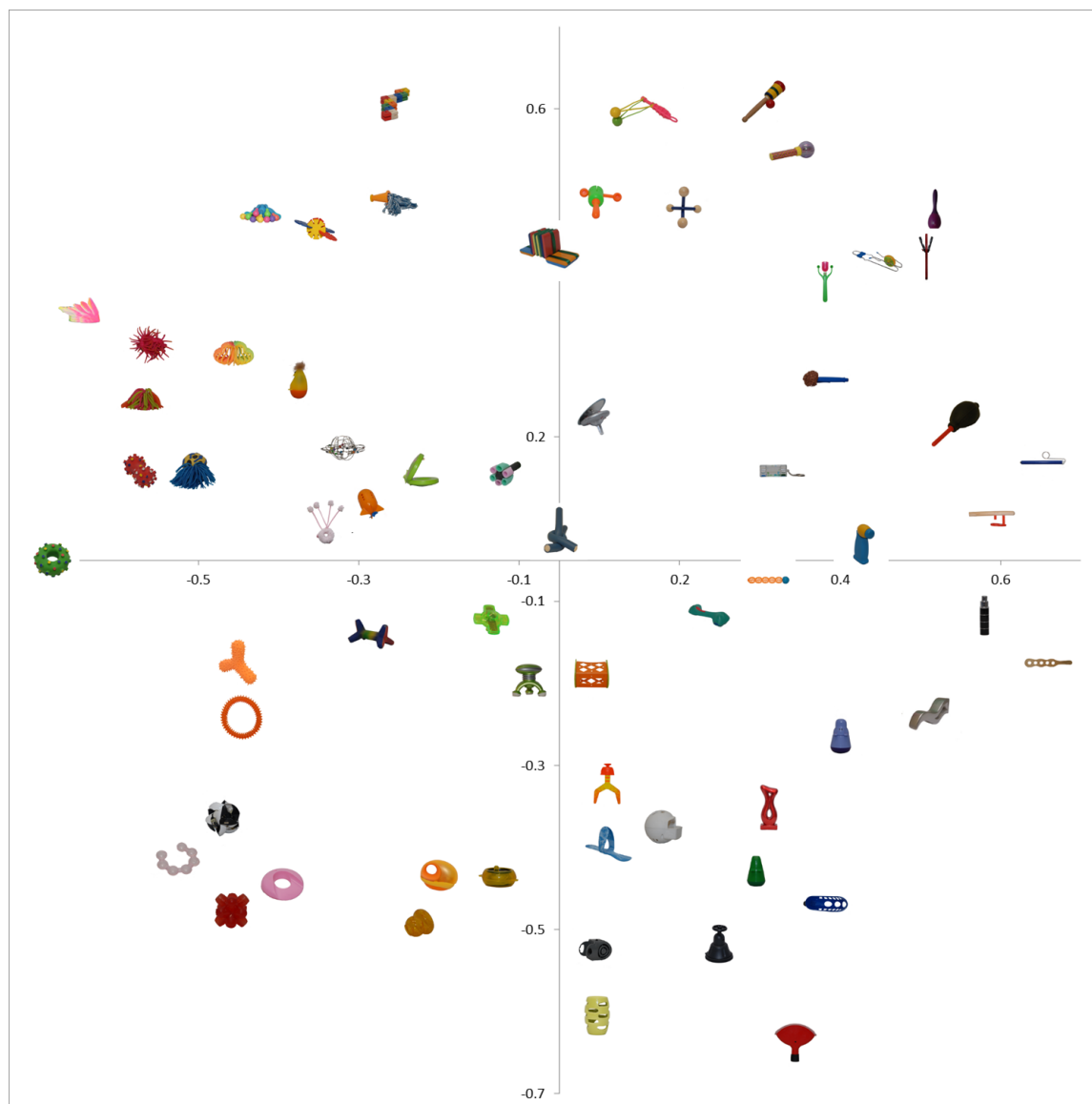


Figure 3. Plotted results of MDS dimensions 1 (X-axis) and 2 (Y-axis), with pictures superimposed, from Experiment 2. The pictures are placed in the image based on their weights on dimension 1 and 2.

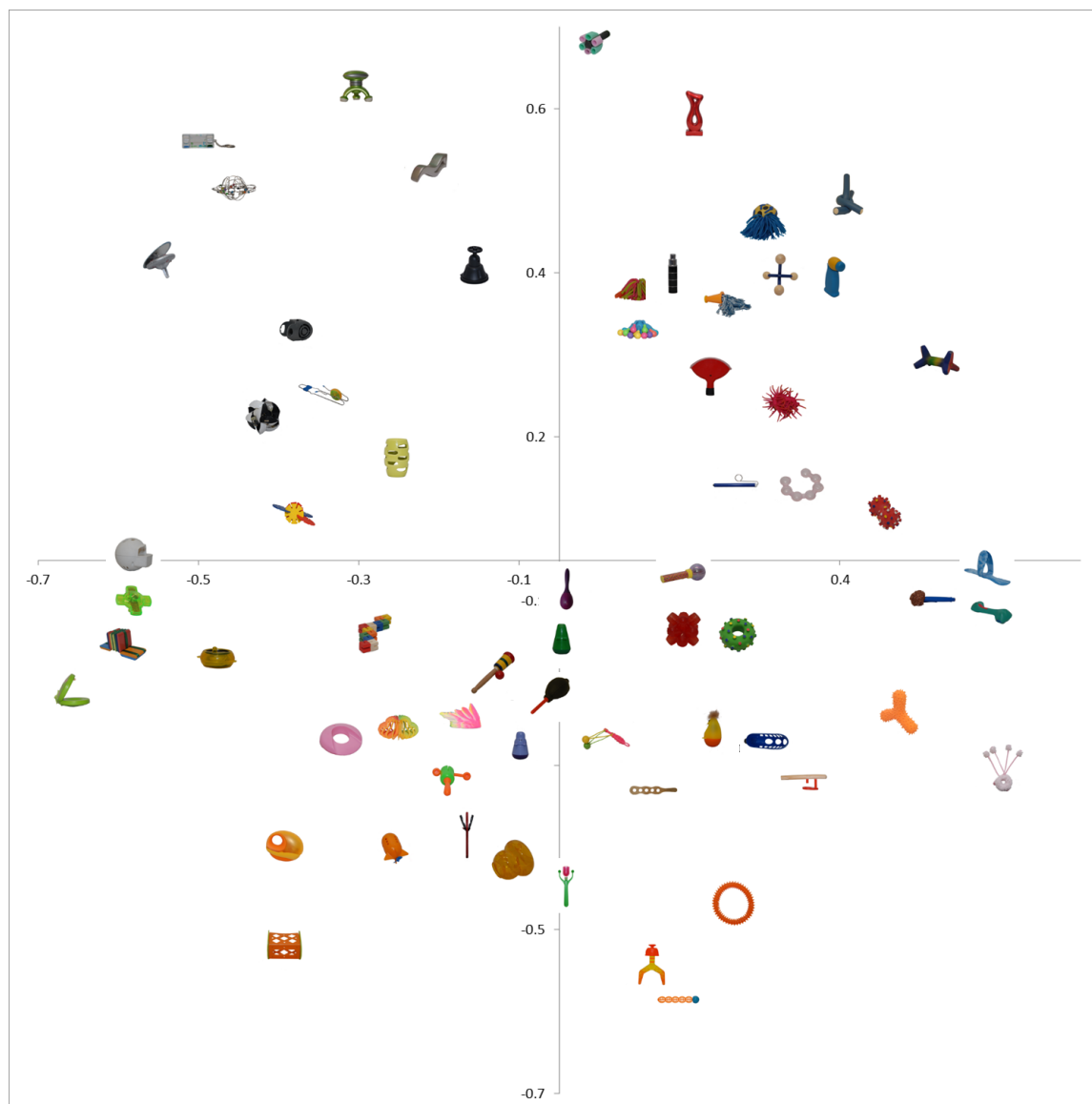


Figure 4. Plotted results of MDS dimensions 3 (X-axis) and 4 (Y-axis), with pictures superimposed (from Experiment 2). The pictures are placed in the image based on their weights on dimension 3 and 4.



Figure 5. Item pairs from Experiment 2, sampled from each classification quartile. The top-left shows the most similar pair of items (items 2052 and 2053), and the bottom-right shows the most dissimilar pair (items 2028 and 2046). The top-right and bottom-left pairs are taken from the top of the second quartile and the bottom of the third quartile, respectively (top-right: items 2030 and 2039; bottom-left: items 2004 and 2031).

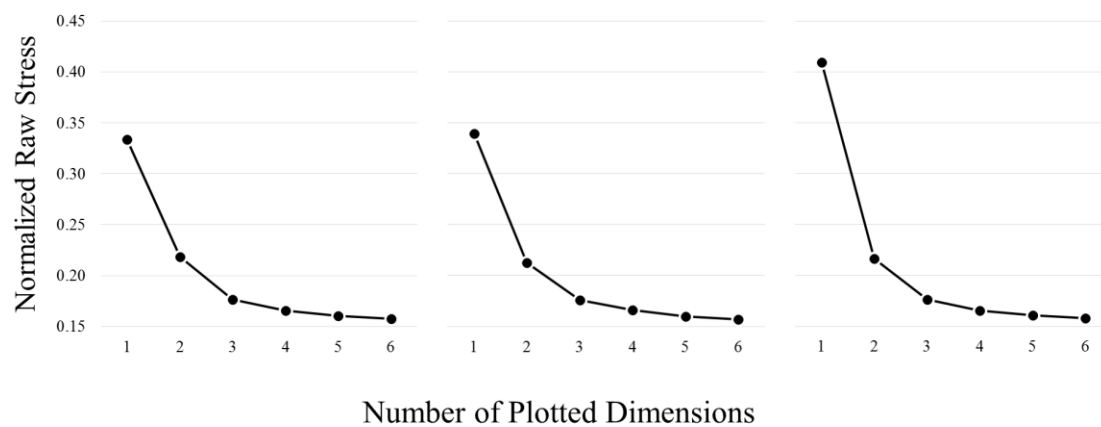


Figure 6. Scree plots for the Simplex (left), Torgerson (middle), and Multiple Random Starts (right) options (from Experiment 3). Stress values are plotted as a function of the dimensionality in which the MDS data were scaled.

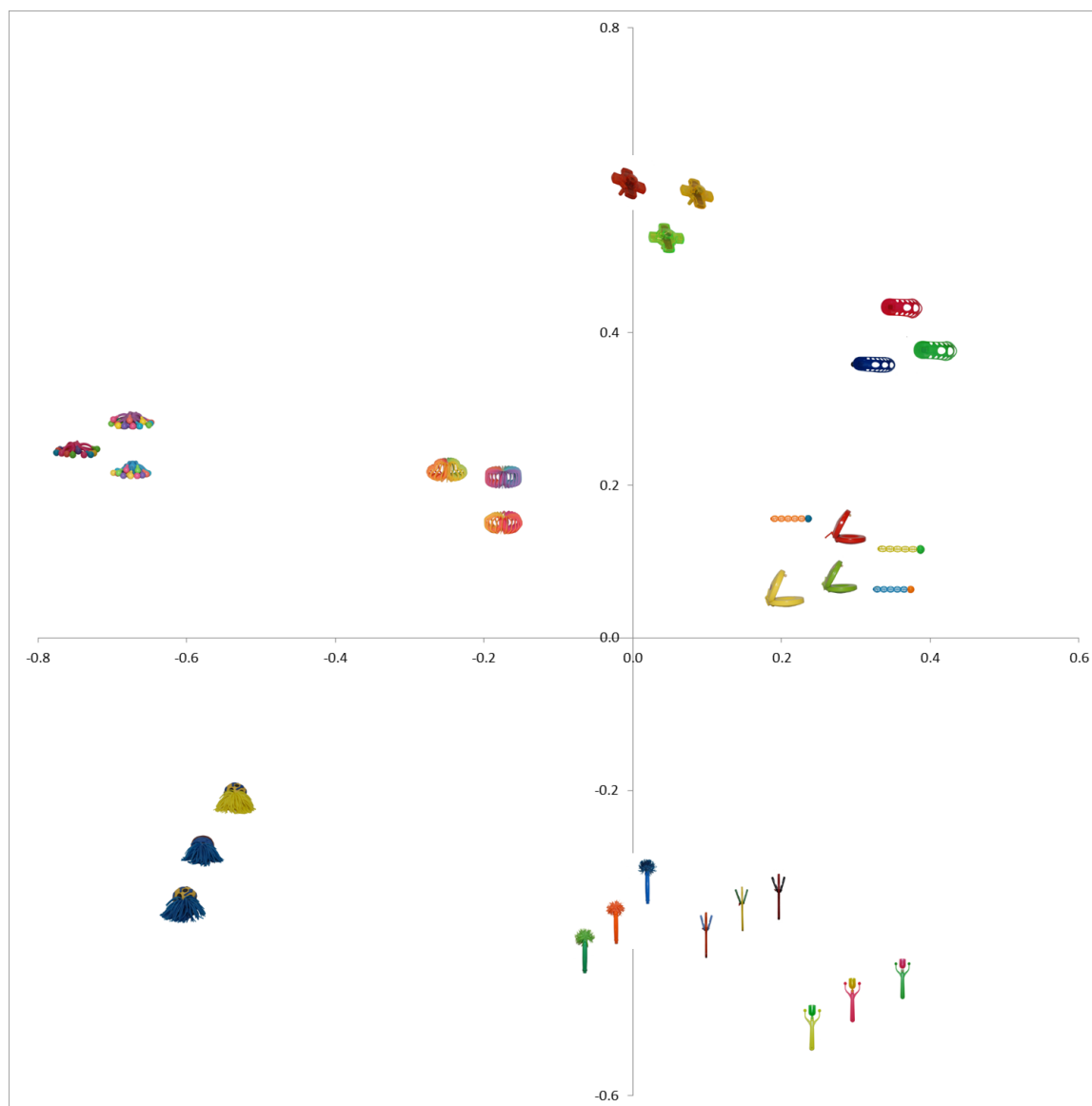


Figure 7. Plotted results of MDS dimensions 1 (X-axis) and 2 (Y-axis), with pictures superimposed (from Experiment 3). The pictures are placed in the image based on their weights on dimension 1 and 2.

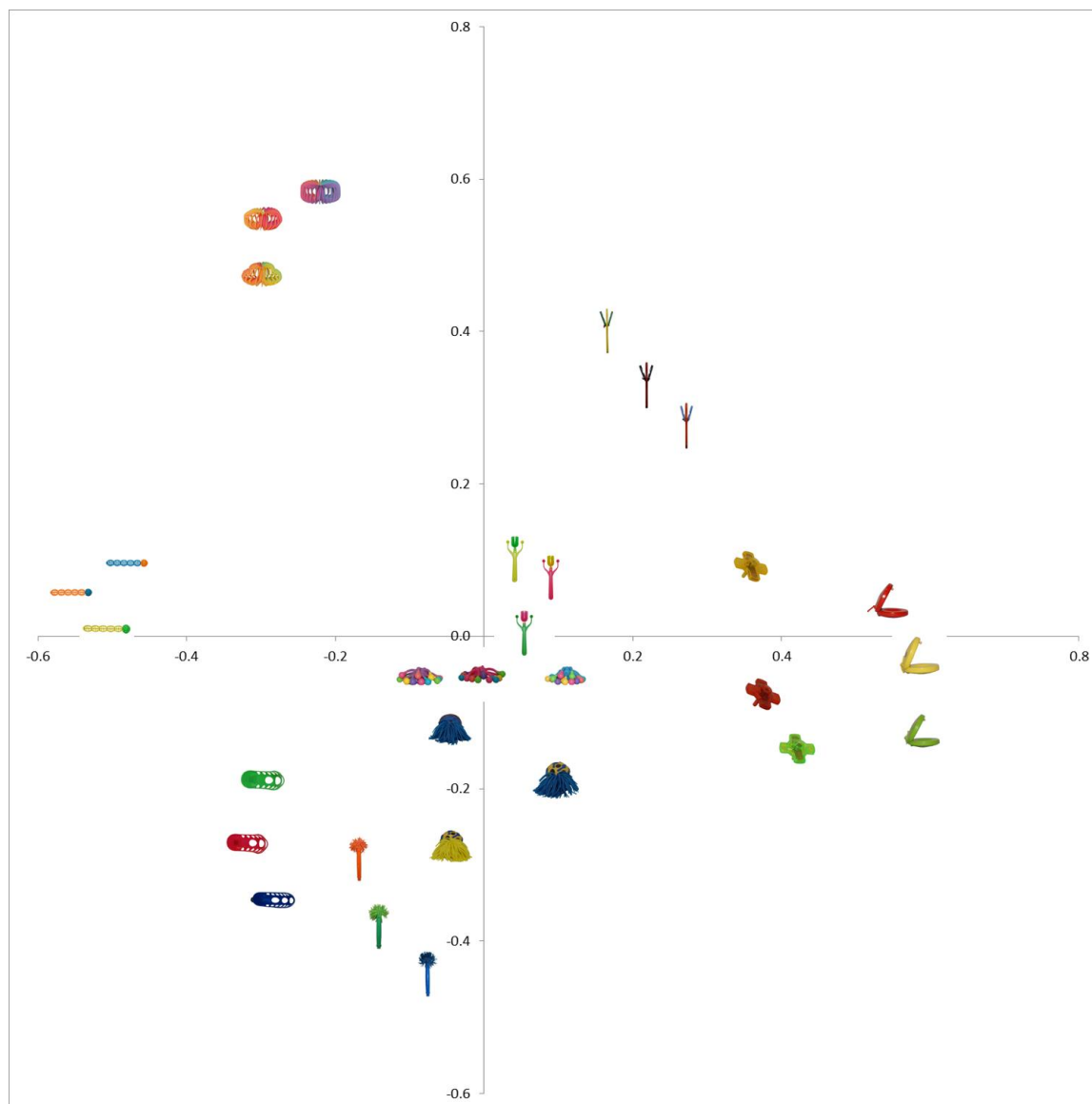


Figure 8. Plotted results of MDS dimensions 3 (X-axis) and 4 (Y-axis), with pictures superimposed (from Experiment 3). The pictures are placed in the image based on their weights on dimension 3 and 4.

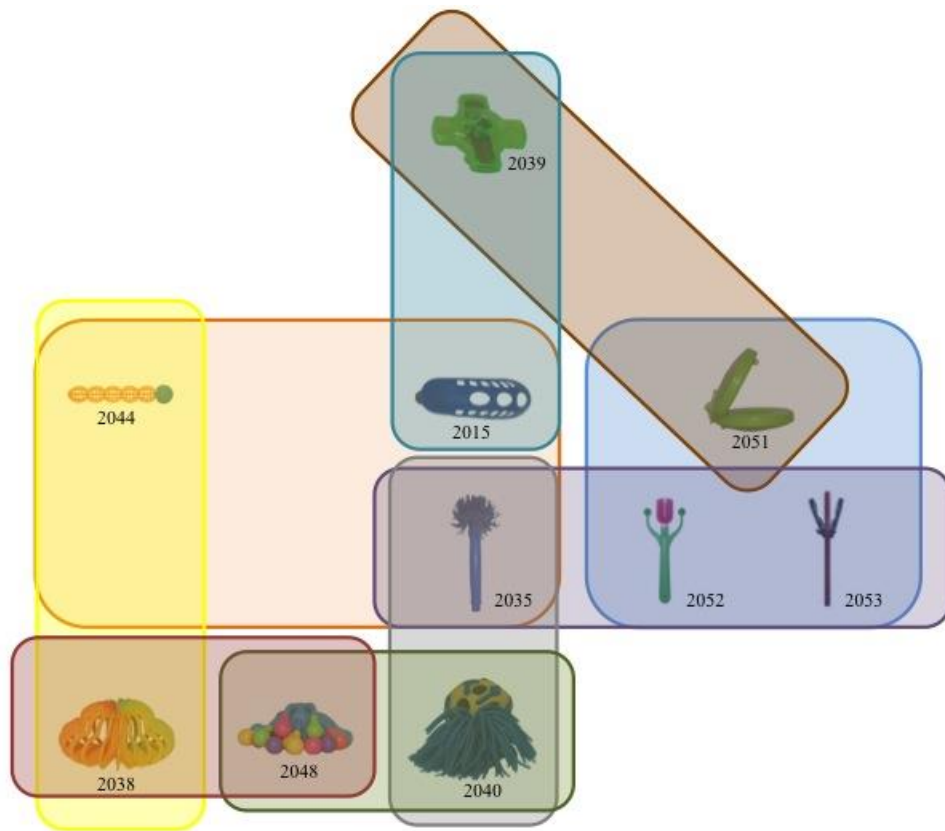


Figure 9. Venn diagram depicting the relationships between the different categories compared in Experiment 3. Three exemplars from each category were compared; only one exemplar per subcategory is shown.